

WHAT IS CLAIMED IS:

1. A method of creating labels for clusters of documents, comprising:
identifying topics associated with the documents in the clusters;
determining whether the topics are associated with at least half of the documents in the clusters;
adding ones of the topics that are associated with at least half of the documents in the clusters to cluster lists; and
forming labels for the clusters from the cluster lists.
2. The method of claim 1, wherein the identifying topics includes:
using a probabilistic Hidden Markov Model to determine the topics.
3. The method of claim 1, wherein the forming labels includes:
ranking the ones of the topics, and
placing the ones of the topics in the labels in ranked order.
4. The method of claim 3, wherein the ranking the ones of the topics includes:
assigning ranks to the ones of the topics based on a number of the documents with which the ones of the topics are associated.

5. The method of claim 1, further comprising:
ranking the ones of the topics based on a number of the documents with which the ones of the topics are associated.
6. The method of claim 5, wherein when a first one of the ones of the topics, as a first topic, is associated with a majority of the documents in one of the clusters and a second one of the ones of the topics, as a second topic, is associated with less than the majority of the documents in the one of the clusters, the first topic is ranked higher than the second topic.
7. The method of claim 5, wherein the ranking the ones of the topics includes:
assigning higher ranks to first ones of the ones of the topics that are associated with larger numbers of the documents than second ones of the ones of the topics that are associated with smaller numbers of the documents.
8. The method of claim 5, wherein the forming labels includes:
sorting the cluster lists based on the rankings of the ones of the topics.
9. A system for generating a label for a cluster of documents, comprising:
means for identifying topics associated with the documents in the cluster;
means for determining whether the topics are associated with at least half of the documents in the cluster; and

means for generating a label for the cluster based on one or more of the topics that are associated with at least half of the documents in the cluster.

10. The system of claim 9, further comprising:

means for ranking the one or more of the topics based on a number of the documents with which the one or more of the topics are associated.

11. The system of claim 10, wherein the means for generating a label includes:

means for sorting the one or more of the topics based on the ranking to form the label for the cluster.

12. A system for creating a label for a cluster of documents, comprising:

logic configured to identify topics associated with the documents in the cluster;

logic configured to determine whether the topics are associated with approximately half or more of the documents in the cluster;

logic configured to rank ones of the topics that are associated with approximately half or more of the documents in the cluster; and

logic configured to generate a label for the cluster using the ones of the topics in ranked order.

13. The system of claim 12, wherein when a first one of the ones of the topics, as a first topic, is associated with a majority of the documents in the cluster and a second one of the

ones of the topics, as a second topic, is associated with less than the majority of the documents in the cluster, the first topic is ranked higher than the second topic.

14. The system of claim 12, wherein the logic configured to rank ones of the topics includes:

logic configured to assign higher ranks to first ones of the ones of the topics that are associated with larger numbers of the documents than second ones of the ones of the topics that are associated with smaller numbers of the documents.

15. The system of claim 12, wherein the logic configured to generate a label includes:
logic configured to sort the ones of the topics based on the rankings of the ones of the topics.

16. A topic detection system, comprising:

a decision engine configured to:

receive a plurality of documents, and

group the documents into a plurality of clusters; and

a label engine configured to:

identify topics associated with the documents in the clusters,

determine whether the topics are associated with at least half of the documents in the clusters, and

form labels for the clusters using ones of the topics that are associated with at least half of the documents in the clusters.

17. The system of claim 16, wherein the label engine is further configured to:
rank the ones of the topics based on a number of the documents with which the ones of the topics are associated.

18. A method for creating labels for clusters of documents, comprising:
identifying topics associated with the documents in the clusters;
determining whether the topics are associated with a predetermined portion of the documents in the clusters; and
generating labels for the clusters using ones of the topics that are associated with approximately half or more of the documents in the clusters.

19. The method of claim 18, wherein the predetermined portion of the documents is equal to approximately half of the documents.